

Modelos lineales generalizados

Javier Fernández-López, Profesor Ayudante Doctor

Unidad de Matemática Aplicada

Departamento de Biodiversidad, Ecología y Evolución, UCM

Bloque 2: Modelos lineales generalizados

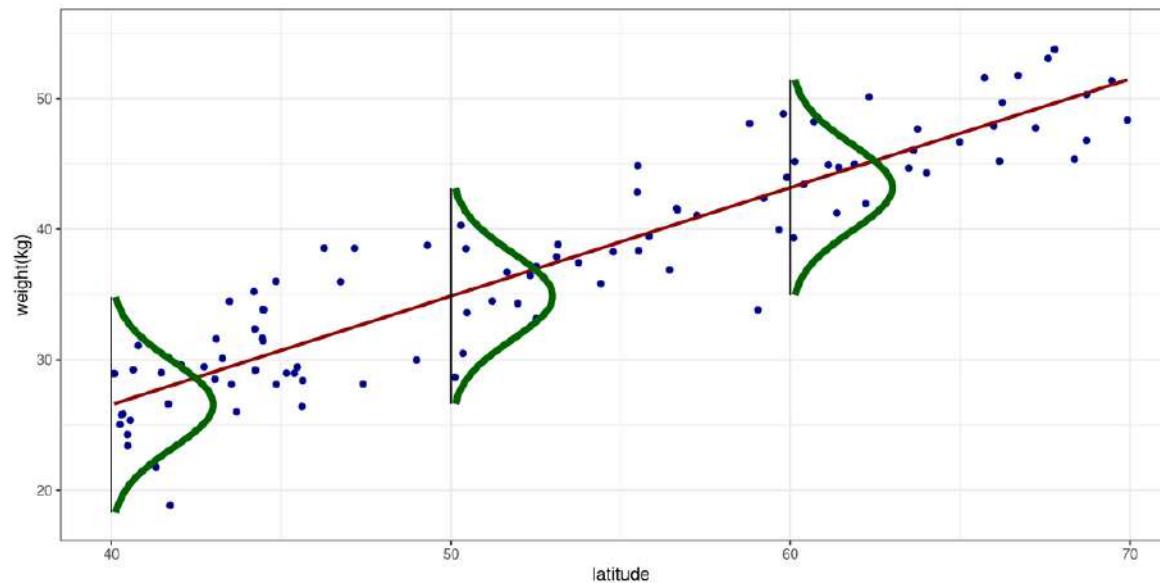
- Extendiendo el modelo lineal: factores categóricos
- Función vínculo: más allá de la distribución normal
- Evaluación y selección de modelos
- Predicciones

Factores de clasificación: predictores categóricos

$$weight_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

Predictor cuantitativo continuo



Factores de clasificación: predictores categóricos

i	<i>concentración</i>	<i>lote</i>
1	27	A
2	35	A
3	34	A
3	42	B
3	46	B
...
n	50	B

Factores de clasificación: predictores categóricos

i	<i>concentración</i>	<i>lote</i>
1	27	A
2	35	A
3	34	A
3	42	B
3	46	B
...
n	50	B

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 \cdot \text{factor}_i$$

Factores de clasificación: predictores categóricos

i	<i>concentración</i>	<i>lote</i>	<i>dummy</i>
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \cdot \text{factor}_i$$

$$\text{factor}_i = \begin{cases} 0, & \text{si el individuo } i \text{ pertenece al primer nivel del factor} \\ 1, & \text{si el individuo } i \text{ pertenece al segundo nivel del factor} \end{cases}$$

Factores de clasificación: predictores categóricos

i	<i>concentración</i>	<i>lote</i>	<i>dummy</i>
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \cdot \text{factor}_i \rightarrow 0$$

$$\text{factor}_i = \begin{cases} 0, & \text{si el individuo } i \text{ pertenece al primer nivel del factor} \\ 1, & \text{si el individuo } i \text{ pertenece al segundo nivel del factor} \end{cases}$$

Factores de clasificación: predictores categóricos

i	<i>concentración</i>	<i>lote</i>	<i>dummy</i>
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \cdot \text{factor}_i$$



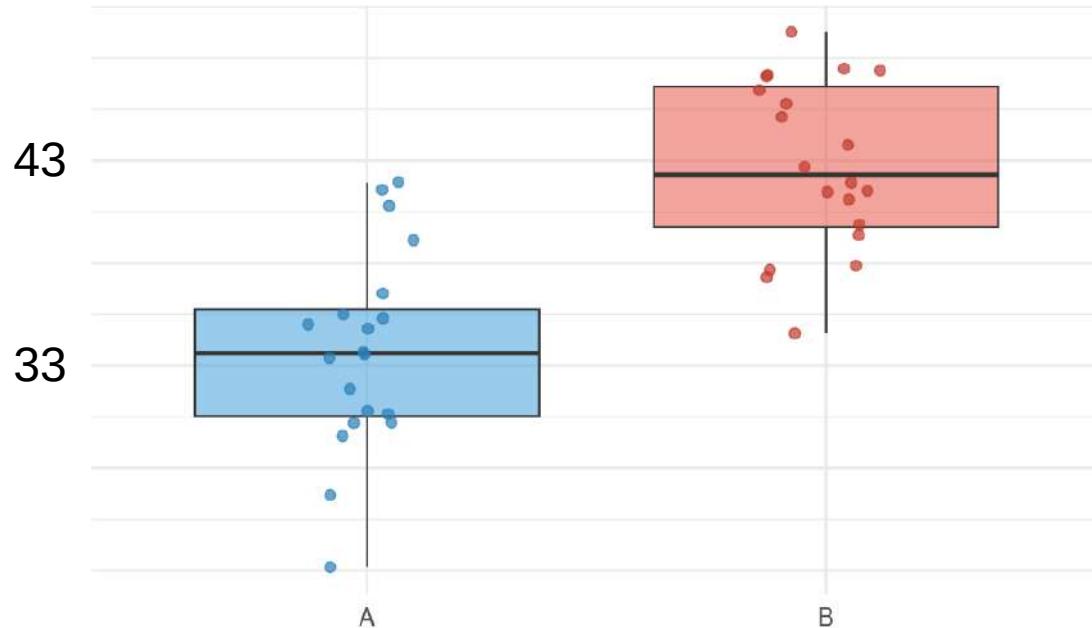
$$\text{factor}_i = \begin{cases} 0, & \text{si el individuo } i \text{ pertenece al primer nivel del factor} \\ 1, & \text{si el individuo } i \text{ pertenece al segundo nivel del factor} \end{cases}$$

Factores de clasificación: predictores categóricos

i	<i>concentración</i>	<i>lote</i>	<i>dummy</i>
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \cdot \text{factor}_i$$



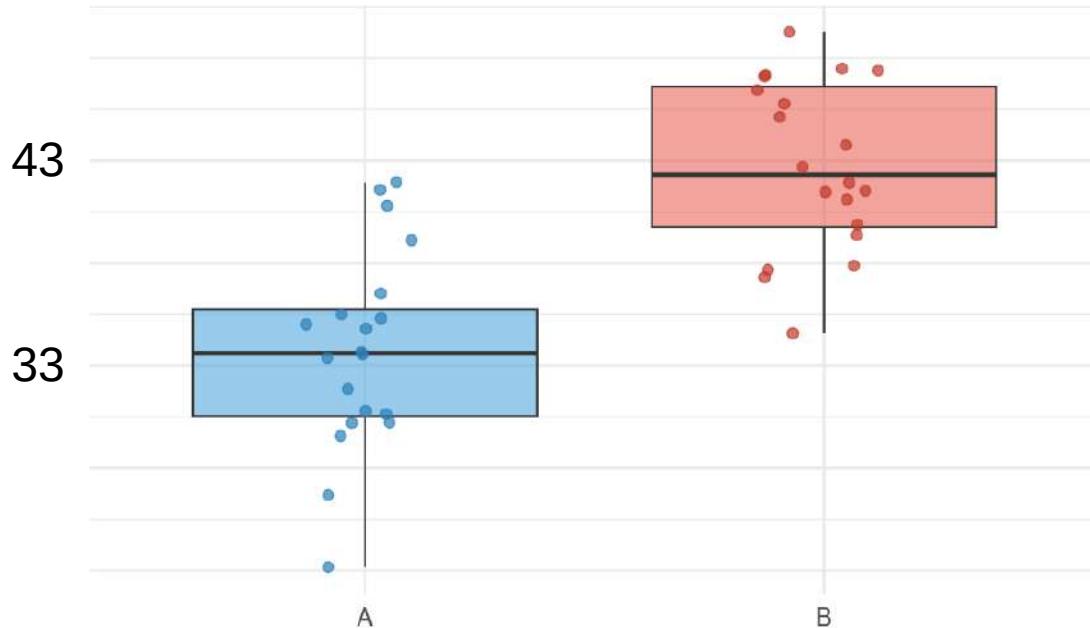
Factores de clasificación: predictores categóricos

i	<i>concentración</i>	<i>lote</i>	<i>dummy</i>
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \cdot \text{factor}_i$$

ANOVA
ANalysis Of VAriance



Factores de clasificación: predictores categóricos

i	concentración	lote	dummy
1	27	A	0
2	35	A	0
3	34	A	0
3	42	B	1
3	46	B	1
...
n	50	B	1

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \cdot \text{factor}_i$$

ANOVA
ANalysis Of VAriance

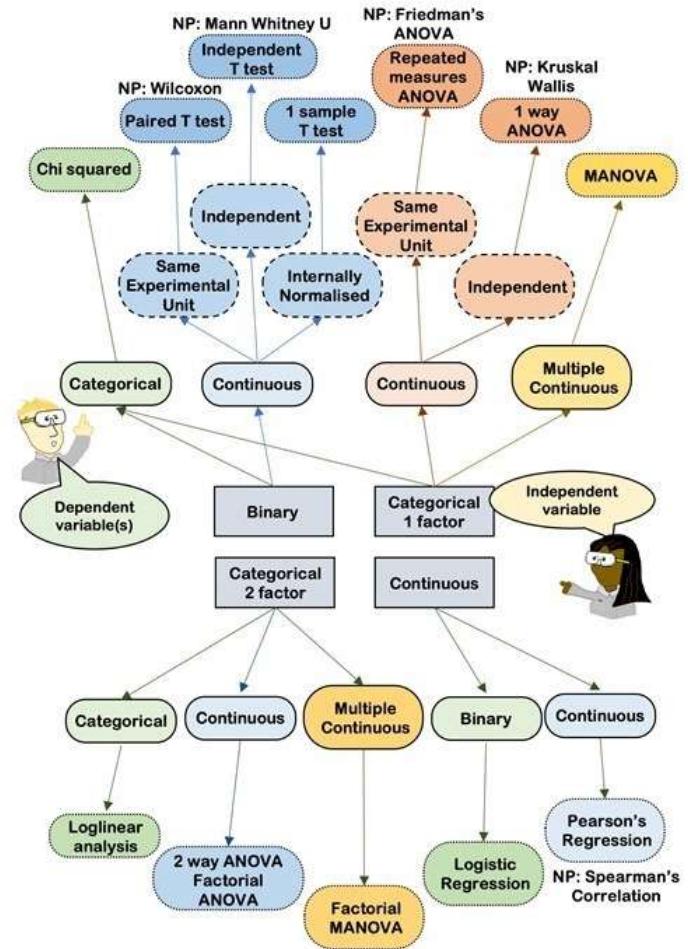
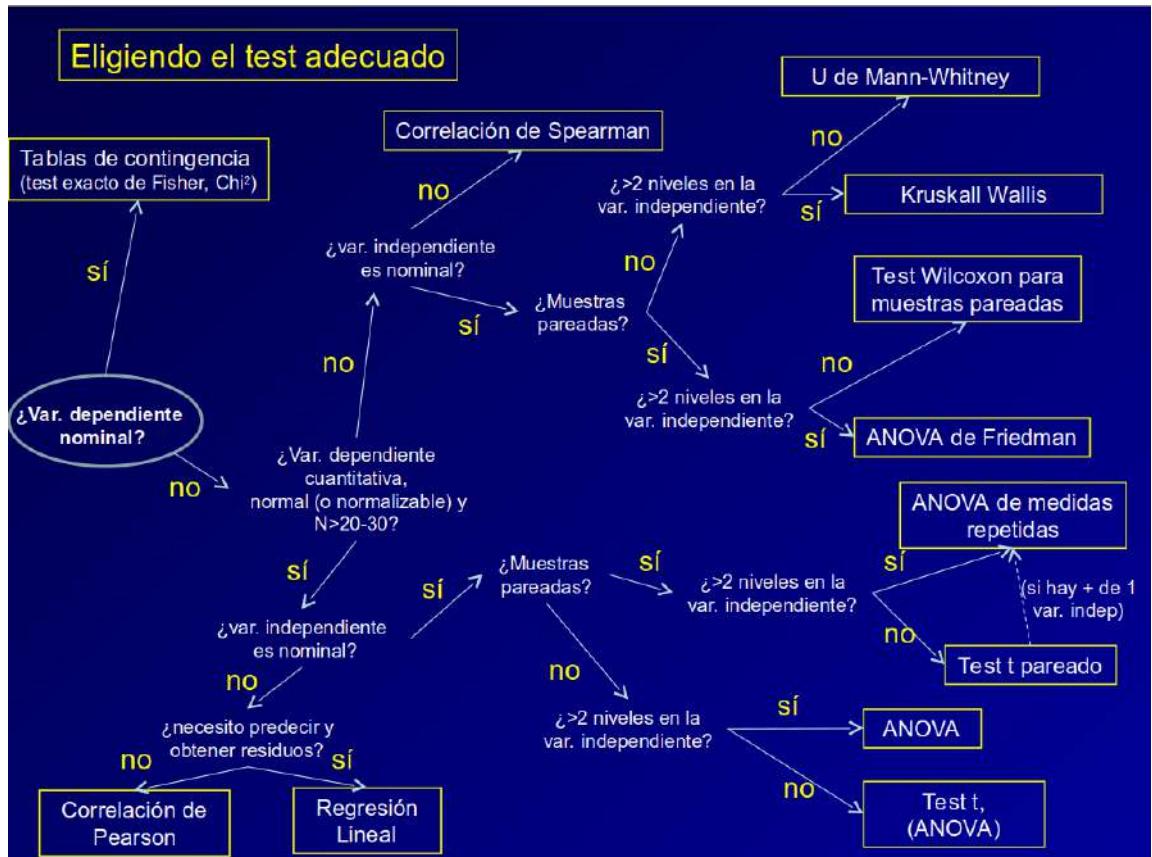
El ANOVA (y similares como ANCOVA, etc.) puede verse como un caso concreto de un modelo general lineal, en el que todos nuestros predictores son categóricos. Por eso, es más cómodo y didáctico entender todos estos tests como casos concretos de un marco de modelización general (GLM).

- 1) Evitamos memorizar numerosos nombres de tests estadísticos.
- 2) Nos obliga a pensar sobre nuestro diseño/variables.
- 3) Crea un marco mental más constructivo.

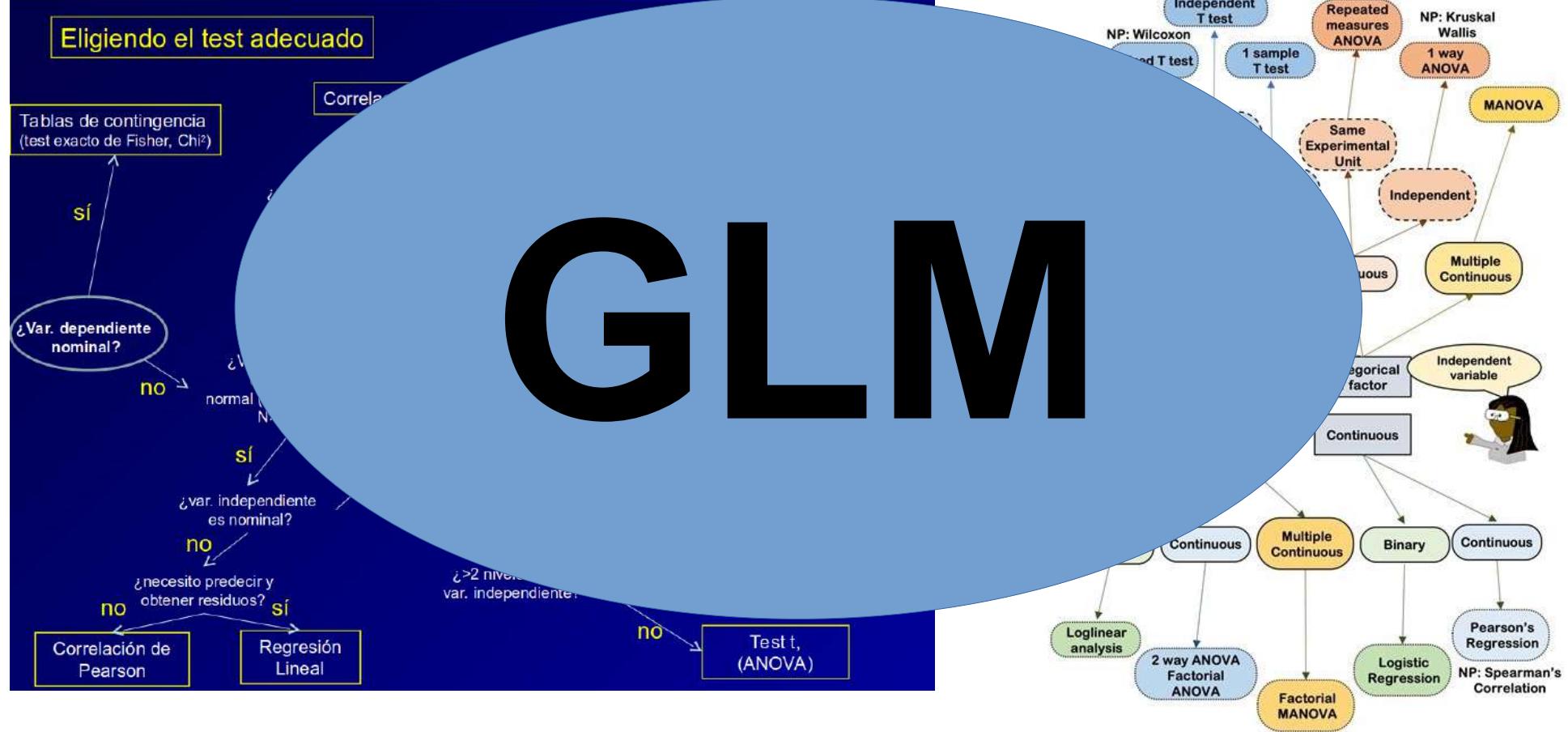
FACTORES CATEGÓRICOS

Modelo	Variables predictoras	Tipo de respuesta	Propósito principal
ANOVA	Categóricas	Continua (normal)	Comparar medias entre grupos
Regresión lineal	Continuas	Continua (normal)	Predecir Y en función de X
ANCOVA	Categóricas + continuas	Continua (normal)	Comparar medias ajustando por covariables
GLM	Categóricas + continuas	Continua (normal)	Marco general que los contiene

FACTORES CATEGÓRICOS



FACTORES CATEGÓRICOS



FACTORES CATEGÓRICOS

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: <https://lindeloev.github.io/tests-as-linear>

Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: $\text{Im}(y \sim 1 + x)$	y is independent of x P: One-sample t-test N: Wilcoxon signed-rank	$\text{Im}(y \sim 1)$ $\text{Im}(\text{signed_rank}(y) \sim 1)$	✓ for $N > 14$	One number (intercept, i.e., the mean) predicts y. - (Same, but it predicts the <i>signed rank</i> of y.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	$\text{Im}(y_2 - y_1 \sim 1)$ $\text{Im}(\text{signed_rank}(y_2 - y_1) \sim 1)$	✓ for $N > 14$	One intercept predicts the pairwise $y_2 - y_1$ differences. - (Same, but it predicts the <i>signed rank</i> of $y_2 - y_1$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	$\text{Im}(y \sim 1 + x)$ $\text{Im}(\text{rank}(y) \sim 1 + \text{rank}(x))$	✓ for $N > 10$	One intercept plus x multiplied by a number (slope) predicts y. - (Same, but with <i>ranked</i> x and y)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	$\text{Im}(y \sim 1 + G_2)^A$ $\text{gls}(y \sim 1 + G_2, \text{weights} = \dots)^A$ $\text{Im}(\text{signed_rank}(y) \sim 1 + G_2)^A$	✓ ✓ for $N > 11$	An intercept for group 1 (plus a difference if group 2) predicts y. - (Same, but with one variance <i>per group</i> instead of one common.) - (Same, but it predicts the <i>signed rank</i> of y.)	
Multiple regression: $\text{Im}(y \sim 1 + x_1 + x_2 + \dots)$	P: One-way ANOVA N: Kruskal-Wallis	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_n)^A$ $\text{Im}(\text{rank}(y) \sim 1 + G_2 + G_3 + \dots + G_n)^A$	✓ for $N > 11$	An intercept for group 1 (plus a difference if $group \neq 1$) predicts y. - (Same, but it predicts the <i>rank</i> of y.)	
	P: One-way ANCOVA	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_n + x)^A$	✓	- (Same, but plus a slope on x.) Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.	
	P: Two-way ANOVA	$\text{Im}(y \sim 1 + G_2 + G_3 + \dots + G_n + S_2 + S_3 + \dots + S_k + G_2 * S_2 + G_3 * S_3 + \dots + G_n * S_k)^A$	✓	Interaction term: changing sex changes the y ~ group parameters. Note: $G_{i,k}$ is an indicator (0 or 1) for each non-intercept levels of the group variable . Similarly for $S_{i,k}$ for sex. The first line (with G_j) is main effect of group, the second (with S_j) for sex and the third is the group * sex interaction. For two levels (e.g. male/female), line 2 would just be " S_2 " and line 3 would be S_2 multiplied with each G_i .	[Coming]
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	✓	Interaction term: (Same as Two-way ANOVA.) Note: Run <code>glm</code> using the following arguments: <code>glm(model, family=poisson())</code> . As <code>linear-model</code> , the Chi-square test is $\log(y) = \log(N) + \log(\alpha) + \log(\beta) + \log(\alpha\beta)$ where α and β are proportions. See more info in the accompanying notebook.	Same as Two-way ANOVA
N: Goodness of fit	chisq.test(y)	$\text{glim}(y \sim 1 + G_2 + G_3 + \dots + G_n, \text{family} = \dots)^A$	✓	(Same as One-way ANOVA and see Chi-Square note.)	1W-ANOVA

Link!

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation $y \sim 1 + x$ is R shorthand for $y = 1 + b + ax$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" **indicator variables** (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. `Im` requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at <https://lindeloev.github.io/tests-as-linear>.

^A See the note to the two-way ANOVA for explanation of the notation.

^B Same model, but with one variance per group: `gls(value ~ 1 + G_2, weights = varIdent(form = ~1|group), method="ML")`.



Caso práctico: fungicidas triazoles en heces de perdiz



Los fungicidas triazoles son compuestos químicos que se aplican habitualmente en semillas de cultivos para prevenir el crecimiento de hongos patógenos de plantas. Sin embargo, cuando las semillas son consumidas por la fauna silvestre, estos compuestos pueden producir efectos crónicos perjudiciales en su salud y desarrollo. Queremos estudiar el efecto de los fungicidas triazoles sobre la condición corporal (peso) en perdices rojas (*Alectoris rufa*). Para ello se han capturado un total de 300 perdices en tres hábitats diferentes (semiurbano, agrícola y monte matorralizado) a las que se les ha sexado y extraído muestras de heces para obtener la concentración de fungicidas triazoles (ngramos de compuesto/gramo de heces).

Caso práctico: fungicidas triazoles en heces de perdiz



- Pensar variables: respuesta y predictores. Tipos de variables
- Construir un modelo desde lo sencillo a lo complejo
- Gráficos diagnóstico y bondad de ajuste

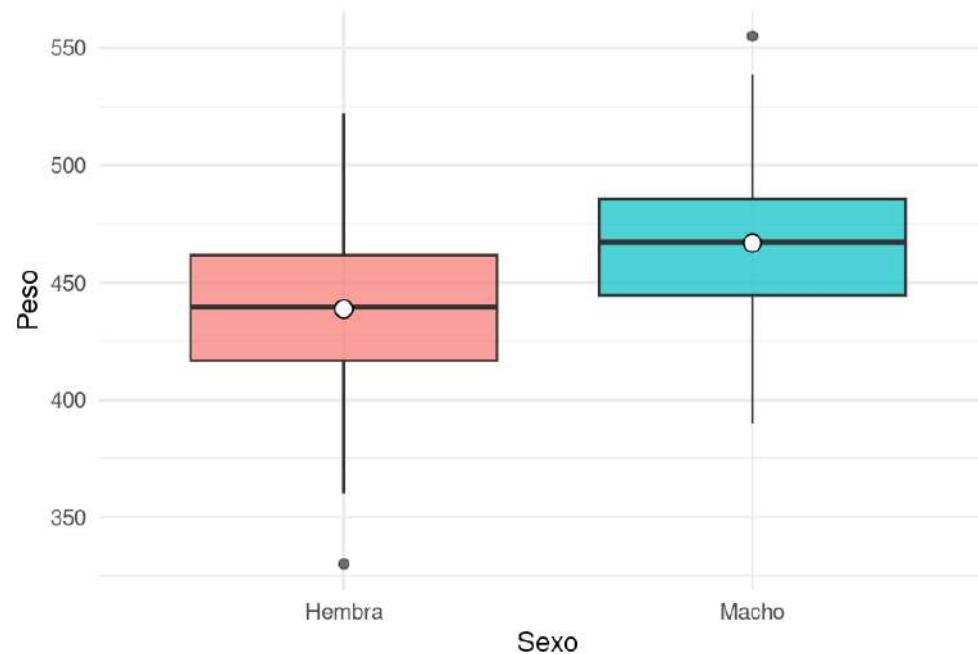
CASO PRÁCTICO

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho})$$

Paso 1: Solo sexo

β_0 = media Hembra; β_1 = (Macho - Hembra)



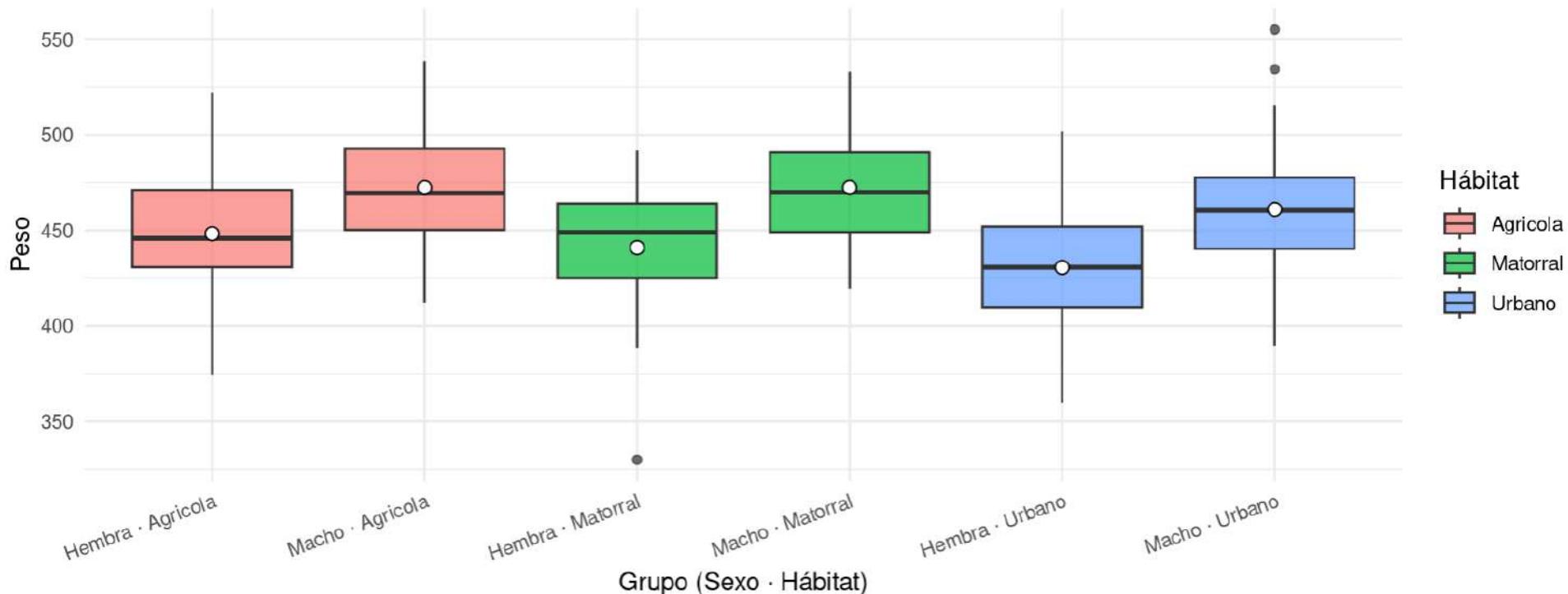
CASO PRÁCTICO

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano})$$

Paso 2: Sexo + Hábitat (6 cajas)

β_0 : Hembra-Agrícola; β_1 : Macho vs Hembra (en Agrícola); β_2 : Matorral vs Agrícola (en Hembra); β_3 : Urbano vs Agrícola (en Hembra)



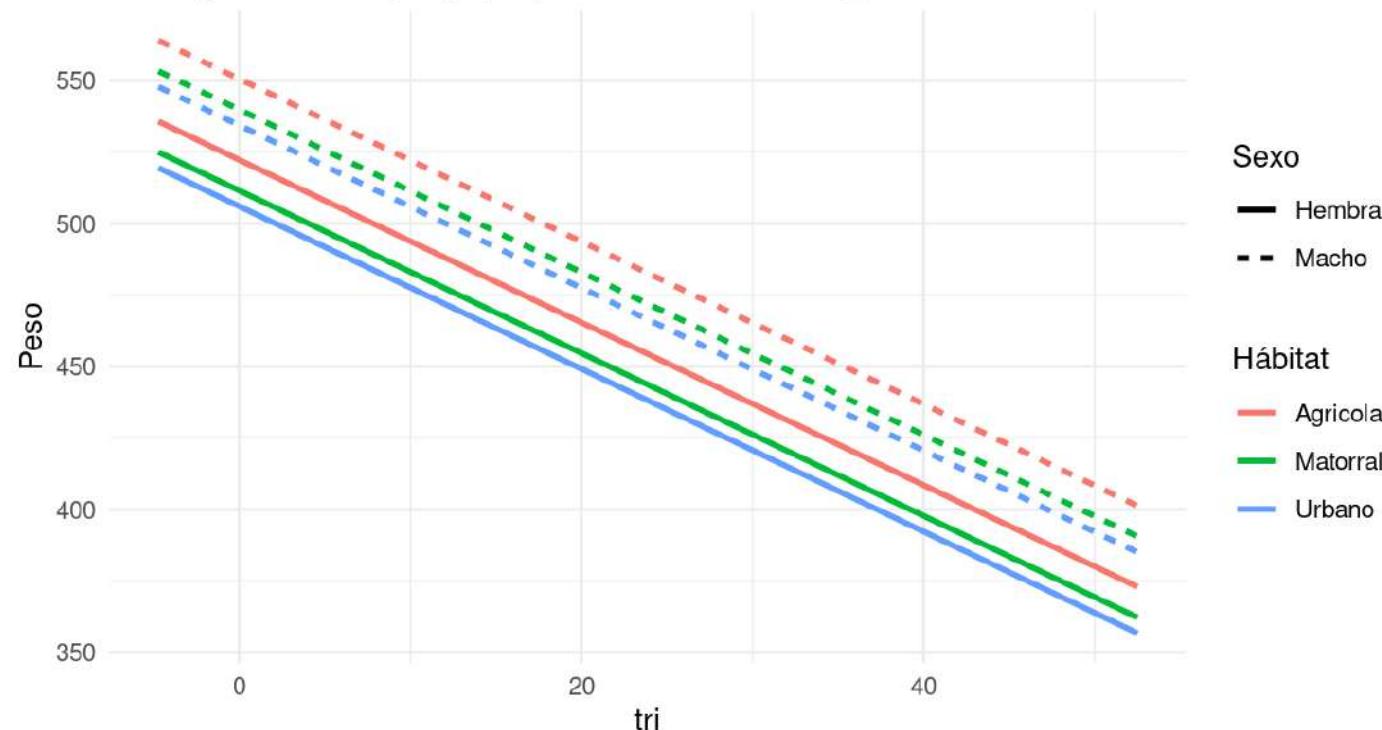
CASO PRÁCTICO

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano}) + \beta_4 \text{tri}_i$$

Paso 3: Misma pendiente para las 6 rectas (modelo: peso ~ sex + hab + tri_c)

Interceptos distintos por grupo; pendiente común de tri_c



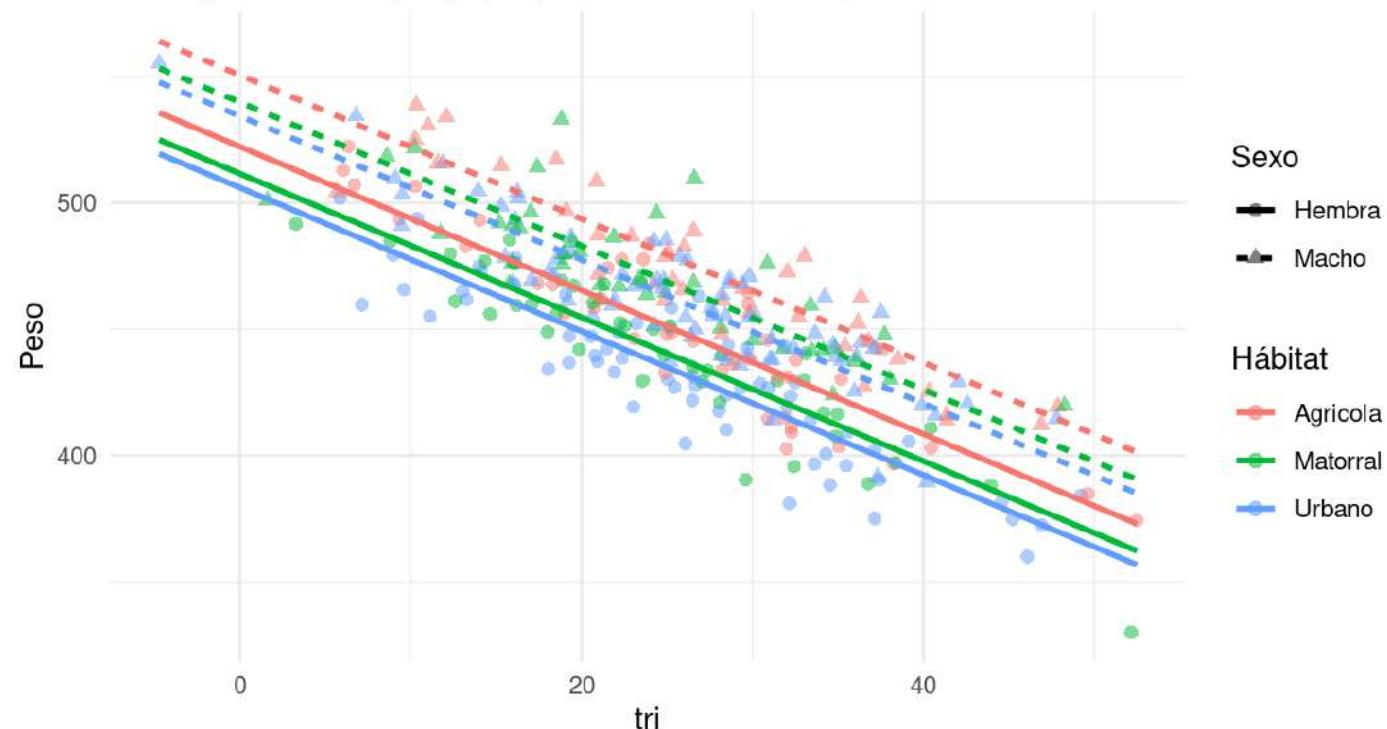
CASO PRÁCTICO

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano}) + \beta_4 \text{tri}_i$$

Paso 3: Misma pendiente para las 6 rectas (modelo: peso ~ sex + hab + tri_c)

Interceptos distintos por grupo; pendiente común de tri_c



PREDICCIONES

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano}) + \beta_4 \text{tri}_i$$

Imaginemos que capturamos una nueva perdiz macho, en hábitat urbano, con una concentración de triazoles de 6 nanogramos por cada gramo de heces. ¿Cuál es el peso predicho para esa perdiz por mi modelo?

$$\mu_{\text{nueva}} =$$

Coefficients:	
	Estimate
(Intercept)	522.22294
sexMacho	28.29805
habMatorral	-10.71749
habUrbano	-16.21739
tri	-2.84291

PREDICCIONES

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano}) + \beta_4 \text{tri}_i$$

Imaginemos que capturamos una nueva perdiz macho, en hábitat urbano, con una concentración de triazoles de 6 nanogramos por cada gramo de heces. ¿Cuál es el peso predicho para esa perdiz por mi modelo?

$$\mu_{\text{nueva}} = 522.22$$

Coefficients:	
	Estimate
(Intercept)	522.22294
sexMacho	28.29805
habMatorral	-10.71749
habUrbano	-16.21739
tri	-2.84291



PREDICCIONES

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano}) + \beta_4 \text{tri}_i$$

Imaginemos que capturamos una nueva perdiz macho, en hábitat urbano, con una concentración de triazoles de 6 nanogramos por cada gramo de heces. ¿Cuál es el peso predicho para esa perdiz por mi modelo?

$$\mu_{\text{nueva}} = 522.22 + (28.3 * 1)$$

Coefficients:	
	Estimate
(Intercept)	522.22294
sexMacho	28.29805
habMatorral	-10.71749
habUrbano	-16.21739
tri	-2.84291



PREDICCIONES

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano}) + \beta_4 \text{tri}_i$$

Imaginemos que capturamos una nueva perdiz macho, en hábitat urbano, con una concentración de triazoles de 6 nanogramos por cada gramo de heces. ¿Cuál es el peso predicho para esa perdiz por mi modelo?

$$\mu_{\text{nueva}} = 522.22 + (28.3 * 1) + (-10.72 * 0)$$

Coefficients:	
	Estimate
(Intercept)	522.22294
sexMacho	28.29805
habMatorral	-10.71749
habUrbano	-16.21739
tri	-2.84291



PREDICCIONES

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \boxed{\beta_3 I(\text{hab}_i = \text{Urbano})} + \beta_4 \text{tri}_i$$

Imaginemos que capturamos una nueva perdiz macho, en hábitat urbano, con una concentración de triazoles de 6 nanogramos por cada gramo de heces. ¿Cuál es el peso predicho para esa perdiz por mi modelo?

$$\mu_{\text{nueva}} = 522.22 + (28.3 * 1) + (-10.72 * 0) + \boxed{(-16.22 * 1)}$$

Coefficients:	
	Estimate
(Intercept)	522.22294
sexMacho	28.29805
habMatorral	-10.71749
habUrbano	-16.21739
tri	-2.84291



PREDICCIONES

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano}) + \boxed{\beta_4 \text{tri}_i}$$

Imaginemos que capturamos una nueva perdiz macho, en hábitat urbano, con una concentración de triazoles de 6 nanogramos por cada gramo de heces. ¿Cuál es el peso predicho para esa perdiz por mi modelo?

$$\mu_{\text{nueva}} = 522.22 + (28.3 * 1) + (-10.72 * 0) + (-16.22 * 1) + \boxed{(-2.84 * 6)}$$

Coefficients:	
	Estimate
(Intercept)	522.22294
sexMacho	28.29805
habMatorral	-10.71749
habUrbano	-16.21739
tri	-2.84291



PREDICCIONES

$$\text{peso}_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 I(\text{sex}_i = \text{Macho}) + \beta_2 I(\text{hab}_i = \text{Matorral}) + \beta_3 I(\text{hab}_i = \text{Urbano}) + \beta_4 \text{tri}_i$$

Imaginemos que capturamos una nueva perdiz macho, en hábitat urbano, con una concentración de triazoles de 6 nanogramos por cada gramo de heces. ¿Cuál es el peso predicho para esa perdiz por mi modelo?

$$\mu_{\text{nueva}} = 522.22 + (28.3 * 1) + (-10.72 * 0) + (-16.22 * 1) + (-2.84 * 6)$$

Coefficients:	Estimate
(Intercept)	522.22294
sexMacho	28.29805
habMatorral	-10.71749
habUrbano	-16.21739
tri	-2.84291

AIC (an information criterion)

Una medida que evalúa qué tan bien nuestro modelo explica nuestros datos equilibrando:

- El **ajuste** del modelo (goodness of fit).
- La **complejidad** del modelo (cuántos parámetros debe estimar)

$$\text{AIC} = 2k - 2 \ln(\hat{L})$$

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.*

In: **Second International Symposium on Information Theory**, edited by B. N. Petrov and F. Csaki, Akademiai Kiado, Budapest, pp. 267–281.

AIC (an information criterion)

Una medida que evalúa qué tan bien nuestro modelo explica nuestros datos equilibrando:

- El **ajuste** del modelo (goodness of fit).
- La **complejidad** del modelo (cuántos parámetros debe estimar)

$$\text{AIC} = \boxed{2k} - 2 \ln(\hat{L})$$

**Nº de parámetros
(complejidad)**

AIC (an information criterion)

Una medida que evalúa qué tan bien nuestro modelo explica nuestros datos equilibrando:

- El **ajuste** del modelo (goodness of fit).
- La **complejidad** del modelo (cuántos parámetros debe estimar)

$$\text{AIC} = 2k - \boxed{2 \ln(\hat{L})}$$

Ajuste

AIC (an information criterion)

Una medida que evalúa qué tan bien nuestro modelo explica nuestros datos equilibrando:

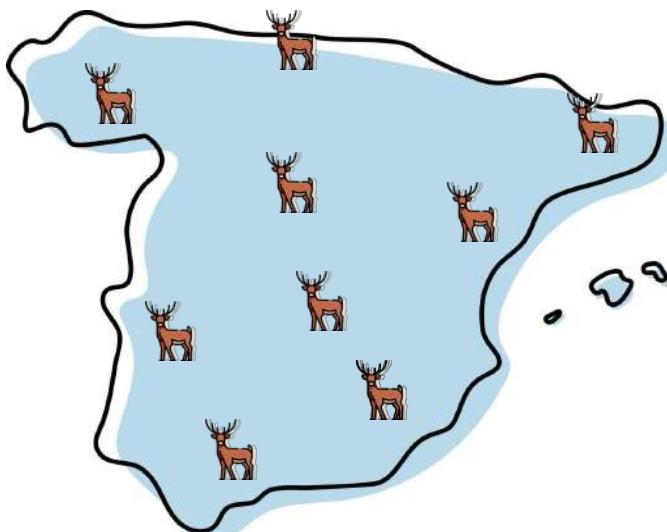
- El **ajuste** del modelo (goodness of fit).
- La **complejidad** del modelo (cuántos parámetros debe estimar).

$$AIC = 2k - 2 \ln(\hat{L})$$

- Menor AIC = mejor modelo (comparativamente).
- Diferencias menores de 2 puntos de AIC se consideran modelos equivalentes.
- Existe una versión corregida para muestras pequeñas (AIC_c). $n/k < 40$ $AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$

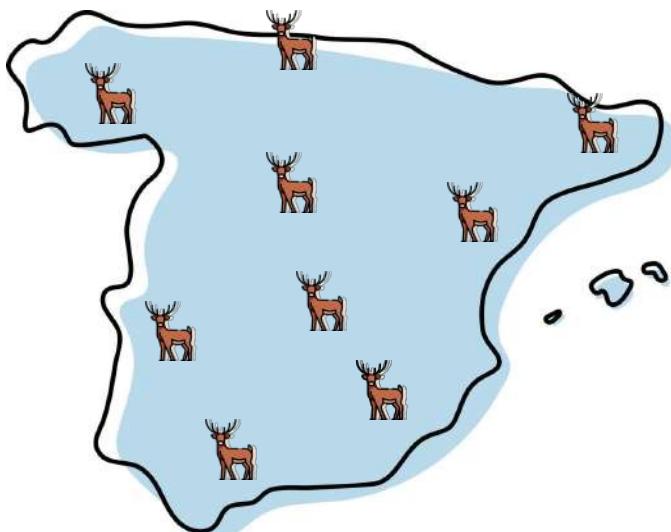
Distribución de la variable respuesta

Hasta ahora hemos utilizado la distribución Normal o Gaussiana para modelizar nuestra variable respuesta, pero no siempre es la más adecuada...



Distribución de la variable respuesta

Hasta ahora hemos utilizado la distribución Normal o Gaussiana para modelizar nuestra variable respuesta, pero no siempre es la más adecuada...

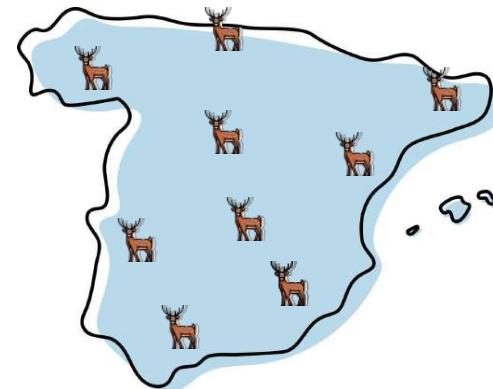


- N.º de agresiones entre gallinas dependiendo del fenotipo.
- Riqueza de sp. de musgos en función de diferentes variables edáficas.
- N.º de germinados/plántulas dependiendo de la influencia de una presa.

Distribución de la variable respuesta

Hemos muestreado 500 cuadrículas de 1x1km contando los ciervos que hemos visto en cada una de ellas... Creemos que el número de ciervos ya a estar relacionado con la temperatura en cada uno de los sitios de muestreo.

Si quisiéramos simular esta situación, ¿qué distribución de probabilidad podríamos utilizar?

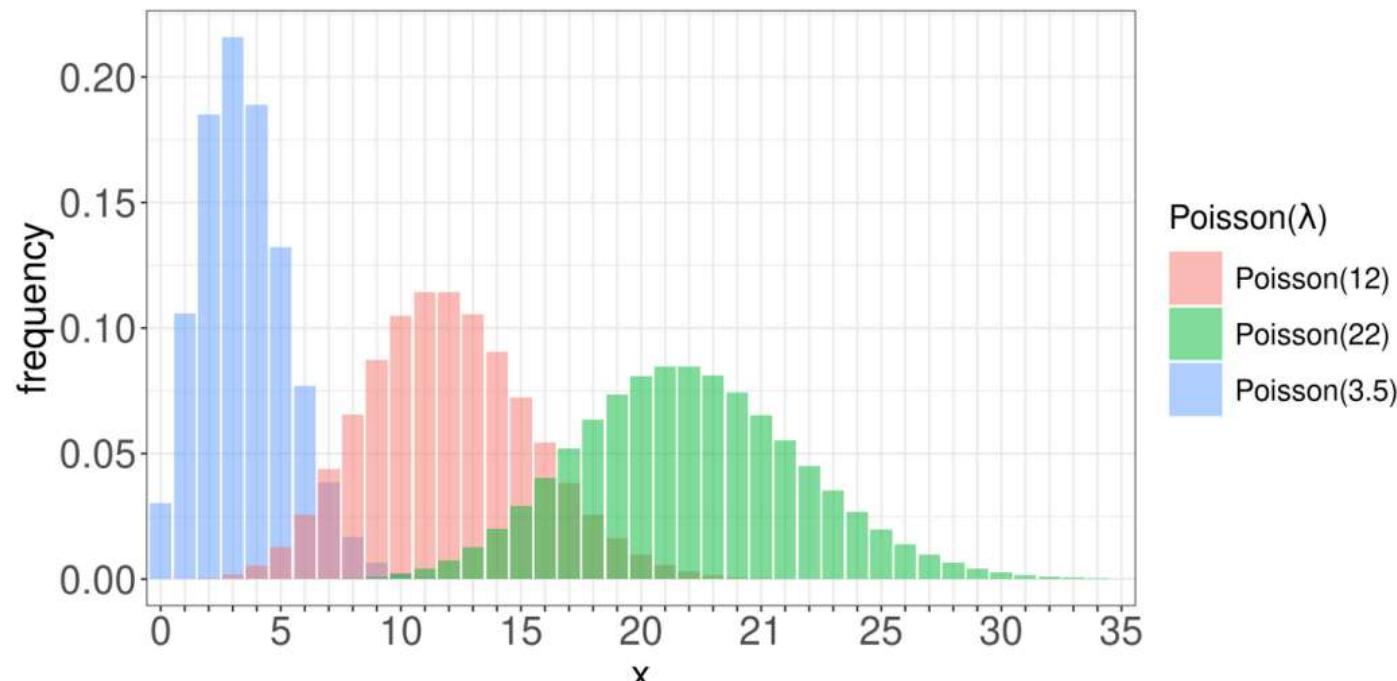


Distribución de la variable respuesta

$$X \sim Poisson(\lambda)$$

Support: $X \in (0, \infty)$ (natural); $\lambda \in (0, \infty)$ (real)

Conteos



Distribución de la variable respuesta: Poisson

$$N_i \sim Poisson(\lambda_i)$$

$$\lambda_i = \beta_0 + \beta_1 Temperatura_i$$

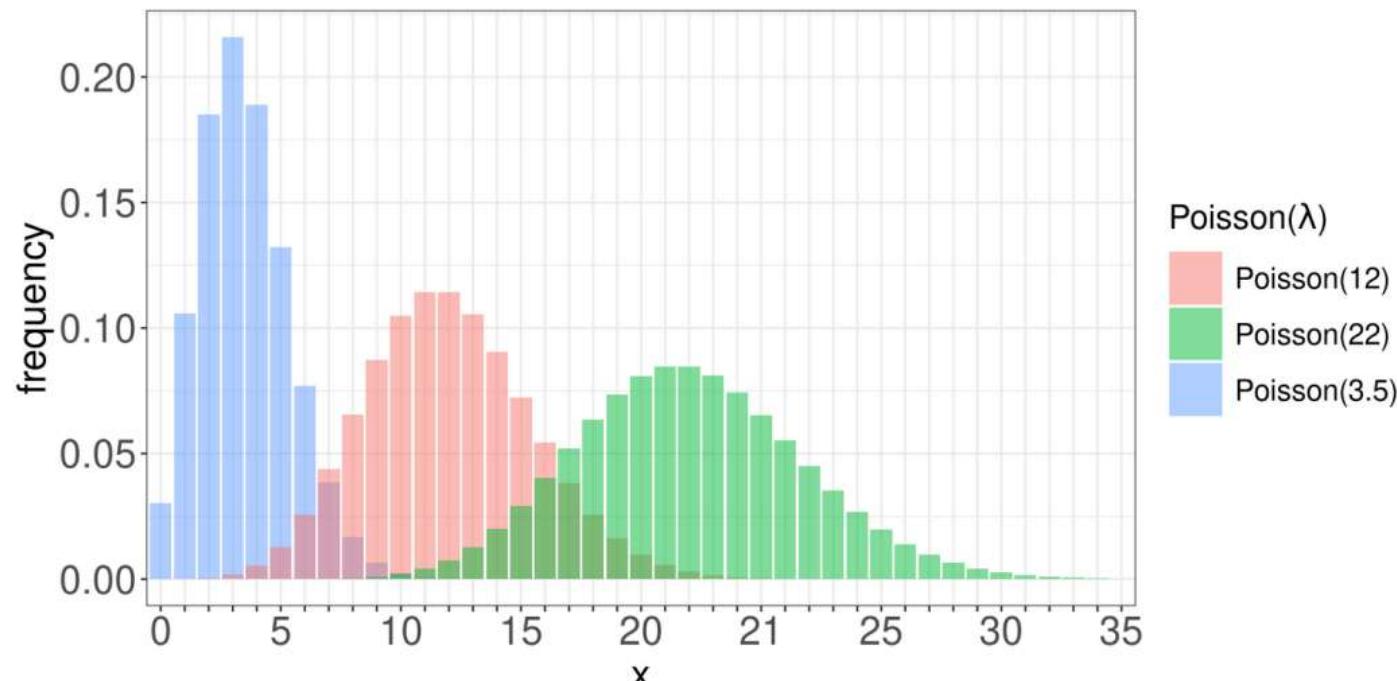


Distribución de la variable respuesta: Poisson

$$X \sim \text{Poisson}(\lambda)$$

Support: $X \in (0, \infty)$ (natural) $\lambda \in (0, \infty)$ (real)

Conteos

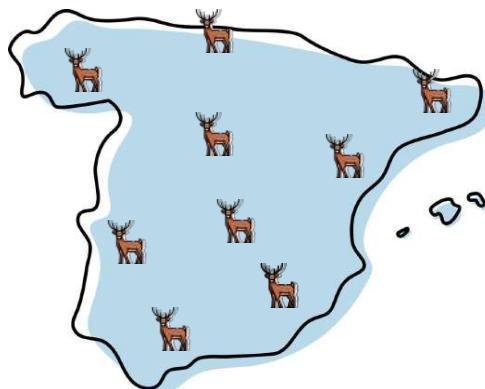


Distribución de la variable respuesta: Poisson

$$N_i \sim Poisson(\lambda_i)$$

$$\lambda_i = \beta_0 + \beta_1 Temperatura_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 Temperatura_i$$



Funciones vínculo

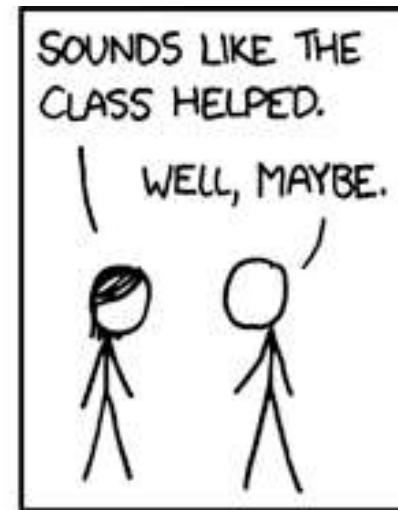
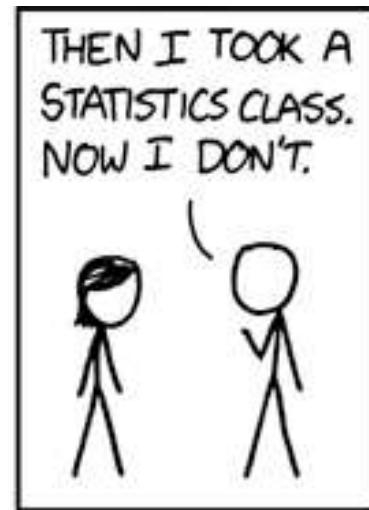
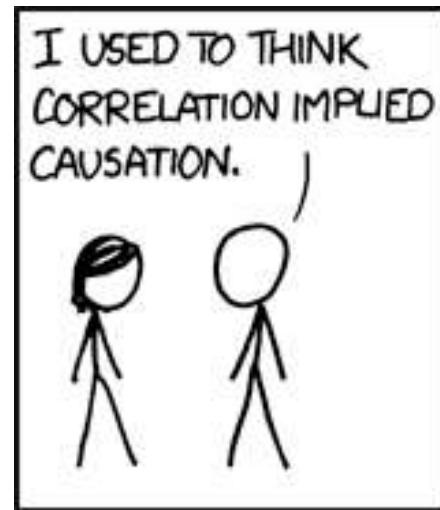
Podemos usar cualquier otra distribución utilizando la misma maquinaria, simplemente se necesitarán funciones vínculo (links) diferentes para unir las covariables con parámetros de cada distribución: log, logit, etc.

Distribución	Links	Fórmula	Inversa
Gaussiana(μ, σ)	Identidad	$\mu = \beta X$	$\mu = \beta X$
Poisson(λ)	Log	$\log(\lambda) = \beta X$	$\lambda = e^{(\beta X)}$
Bernoulli(p)	Logit, probit, cloglog	$\log\left(\frac{p}{1-p}\right) = \beta X$	$p = \frac{e^{(\beta X)}}{1+e^{(\beta X)}}$
Binomial(n,p)	Logit, probit, cloglog	$\log\left(\frac{p}{1-p}\right) = \beta X$	$p = \frac{e^{(\beta X)}}{1+e^{(\beta X)}}$

Con vuestros datos...

- Estudia tu variable respuesta. ¿Qué tipo es? Continua, categórica, admite decimales, valores negativos... ¿Puedo hacer un histograma con ella? ¿Qué pinta tiene?
- ¿Que distribución de probabilidad es la más conveniente para mi variable respuesta?
- Si es posible, elige dos predictores, uno continuo y otro categórico (en el caso de no disponer de esos tipos, elige dos predictores cualesquiera). Examina los predictores (histogramas, boxplot, gráficos de dispersión frente a la variable respuesta).
- Ajusta 4 modelos: el modelo nulo (~ 1), uno con cada variable y uno con ambas.
- Estudia: gráficos diagnóstico, p-valores, coeficientes para cada predictor, y devianza.

CORRELACIÓN EN LOS PREDICTORES



CORRELACIÓN EN LOS PREDICTORES

- Recordamos que lo que hacemos al modelizar es intentar relacionar una información (datos) con una serie de predictores, “repartiendo” el poder explicativo o predictivo entre cada uno de ellos.
- Sin embargo, si dos predictores contienen la misma información, es lógico que el modelo no tenga siempre claro cómo repartir ese poder explicativo... ya que ambos dicen lo mismo!
- La correlación entre predictores puede producir inestabilidad en las estimas de los coeficientes del modelo o un aumento de la incertidumbre en esas estimas.
- No es algo que “tengamos que eliminar”, sino algo que debemos tener en cuenta a la hora de realizar inferencias a partir de nuestro modelo. Suele afectar a las

VIF

El VIF (Variance Inflation Factor) es una forma sencilla y bonita de determinar si la información de un predictor ya está contenida en el conjunto del resto de predictores.

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

Cuanto mayor es el VIF, más redundante es nuestro predictor entre el conjunto de todos los predictores del modelo. Hay diferentes umbrales, pero VIF mayores de 4 suelen considerarse altos.

MODELIZACIÓN DE RELACIONES NO LINEALES

- Una de las asunciones de los modelos lineales era, precisamente, que la relación entre los predictores y la variable respuesta era lineal. Sin embargo, en la mayoría de situaciones esto no es completamente cierto.
- Se pueden incluir términos polinomiales (predictor + predictor² + predictor³) para conseguir un efecto no lineal de los predictores.
- Otra opción es utilizar *splines* mediante General Additive Models.